



Information Research – Vol. 31 No. iConf (2026)

An LLM-powered framework for hierarchical topic discovery in LLM research

Young Soo Ko and Xuan Lu

DOI: <https://doi.org/10.47989/ir31iConf64158>

Abstract

Introduction. The advancement of large language models (LLMs) has led to a rapidly expanding body of research, making systematic mapping of the research landscape critical for informed policy and resource allocation. We present an LLM-powered framework constructing hierarchical topics to reveal how research emerges and reorganizes.

Method. We gather Web of Science (WoS) titles/abstracts, extract topics via GPTopic, and prompt an LLM to form a hierarchy. We integrate quantitative and qualitative analyses, building a bipartite network linking domains and methods.

Analysis. The hierarchical topics provide an overview of the LLM research landscape. Trends are derived from monthly aggregations of publications, offering insights into how research topics evolve and shift in the future.

Results. AI/ML led overall output, peaking in August 2024, with multimodal learning, systems efficiency, and cybersecurity emerging as key growth engines. Medical LLM research evolved from exploration (2023) to workflow integration (2024), reaching specialized deployment focused on fairness and guidelines by 2025 across medical imaging topics. We successfully forecast two research directions that account for 83.75% of publications in a specific area over the next three months.

Conclusion. The framework can effectively generate hierarchical topics for LLM-related research for downstream analysis and can be generalized to other domains.

Introduction

Understanding how research topics emerge, shift, and reorganize in large-scale literature is essential for research investment, policy design, and practical deployment. The development of major innovations could influence research directions across disciplines, contribute to scientific breakthroughs, and even bring about foundational changes. A notable example is Deep Learning (LeCun et al., 2015), which fundamentally transformed research in domains such as computer vision, natural language processing, and medicine. A more recent innovation, Large Language Models (LLMs), has been increasingly studied not only in the AI community but also in various fields such as physics (Barman et al., 2025; Pan et al., 2025), biology (Jumper et al., 2021; Mendoza-Revilla et al., 2024), and social science (Abdurahman et al., 2025; Xu et al., 2024). Specifically, the annual scholarly output related to LLMs has grown by an average of 220.74% per year from 2019 to 2024 (Cotfas et al., 2025). How are LLMs being adopted across different fields? Within a given field, in what ways do LLMs benefit and challenge research? What are the next major questions that could be addressed with LLMs in this field? At the same time, how are LLMs constrained by domain-specific problems, and how might the AI community improve them? Such questions naturally arise with major innovations like LLMs and addressing them can provide a comprehensive overview as well as insightful implications for researchers. Existing works (Patil & Gudivada, 2024; Zhao et al., 2023) have reviewed the evolutions and applications of LLMs, mostly from a technical perspective. Our work, however, aims to address these questions through an LLM-powered framework that can automatically summarize research topics from literature, organize them in a hierarchical structure, and provide time-series analysis with prediction capability.

Our framework is based on topic modeling, an effective approach for extracting latent themes from large corpora and tracing their temporal dynamics. Topic modeling has evolved over time: statistical models such as LDA (Blei et al., 2003) and pLSA (Hofmann, 1999) dominated the early 2000s; context-embedding approaches like BERTopic (Grootendorst, 2022) and Top2Vec (Angelov, 2020) became prevalent in the 2020s; and since 2023, LLM-prompt-based topic modeling such as TopicGPT (Pham et al., 2023) and GPTopic (Reuter et al., 2024) has rapidly proliferated. This approach provides natural-language labels and descriptions for topics, making it more user-friendly and easier to interpret (Pham et al., 2023). However, because topic modeling often operates over broad thematic spaces, it can be challenging to control the granularity of topics (Mu et al., 2024), and fine-grained but meaningful topics may be missed. While GPTopic allows users to adjust topic granularity by interactively splitting and merging topics via prompts, but it becomes cumbersome when the number of topics is large, as each extracted topic must be prompted individually for splitting or merging. Similarly, TopicGPT can generate subtopics from initial topics; however, it does not support deeper refinement, thereby limiting the granularity of the final topics.

To address these challenges, this study proposes an LLM-powered framework to generate multi-level hierarchical topics from literature, building on topics extracted via an existing topic model, and to facilitate a comprehensive understanding of research topic evolution. To uncover research topics related to LLMs, we first collect publications with the keyword ‘LLM’ from the Web of Science Core Collection and obtain their titles and abstracts. We then use GPTopic to generate topics, each with a label and a description. Using these initial topics, we prompt an LLM to generate main themes and apply a top-down process to derive two levels of subthemes. The three levels of themes and the initial topics together constitute a four-level hierarchical theme structure. With this hierarchical structure, topic trends at different levels can be identified to reveal how research foci evolve, thereby providing insights at varying levels of granularity. In addition, to uncover the dynamics of specific topics over time, our framework extracts key terms from publications within a topic and analyzes shifts in the topic through prompting. Furthermore, by identifying the domains and research methods for each publication using the extracted key terms, our framework enables the construction of a bipartite network between domains and methods to illustrate how

LLM-related research methods are adopted across different domains. Our contributions include the following:

- We propose an LLM-powered framework that automatically generates topics from literature and organizes them into a hierarchical structure, facilitating interpretation and downstream analysis at varying levels of granularity.
- Using the proposed framework, we provide an overview of LLM-related research topics with a hierarchical structure and showcase the monthly trends of themes at different levels, offering valuable insights for researchers, practitioners, and decision-makers.
- We select specific topics (two on image and data processing) for a case study of in-topic dynamics. Using key terms extracted from publications within these topics, we conduct a detailed qualitative analysis to examine how the topics have shifted over time and demonstrate the capability of our framework to forecast future developments in these areas.
- We construct a bipartite network between domains and methods within the selected topics. The network analysis reveals both convergence hotspots and latent gaps across fields, providing actionable cues for downstream experimental design and data collection strategies.

Related work

Reviews on LLM research

Large language models (LLMs) have transformed natural language processing (NLP) and related domains, necessitating systematic reviews to consolidate their rapid evolution. Recent literature has organized the LLM development pipeline into key components: architecture design, pretraining and fine-tuning methodologies, efficiency techniques, long-context handling, multimodal and agent-augmented inference, benchmarking frameworks, and safety and ethical considerations (Chang et al., 2024; Cotfas et al., 2025; Naveed et al., 2025). Simultaneously, domain-specific applications in fields such as healthcare and education have garnered focused attention (Dave et al., 2024; Jeon & Lee, 2023).

Chang et al. (2024) organize LLM evaluation along three axes—what to evaluate, where to evaluate, and how to evaluate—thereby providing a structured spectrum that spans general NLP tasks, reasoning, healthcare, ethics, education, science, and agentic applications. Cotfas et al. (2025) employ a bibliometric analysis based on the Web of Science database to map the macro-level landscape of LLM research, visualizing country and journal distributions and summarizing highly cited works to reveal global trends at a glance. Naveed et al. (2025) present an integrated view of the full LLM lifecycle—architecture, training, efficiency, context extension, multimodal/agent capabilities, benchmarking, and safety/ethics—linking state-of-the-art methods (e.g., preference alignment, DPO/RAFT/RRHF, context extension, and model compression) to application domains such as healthcare, robotics, and security to offer a practical comparative framework. Jeon and Lee (2023) illustrate real-world feasibility through case-based analyses, articulating use contexts in education and healthcare, expected benefits (productivity, learning support, decision assistance), and operational constraints (cost, latency, quality control). Dave et al. (2024) provide a concise overview of ChatGPT in healthcare, organizing the discussion around applications, benefits, limitations, ethics, and future directions.

Despite their contributions, many LLM review articles suffer from significant limitations. A common issue is the temporal lag—often exceeding one year—between the last search date and publication, which hinders the inclusion of cutting-edge developments (Elliott et al., 2014). Additionally, most reviews rely on narrative synthesis, offering limited insight into fine-grained topical structures or the dynamic evolution of research topics over time. These gaps restrict the ability to capture emerging trends and reorganizations within the LLM research landscape. To address these gaps, the present study proposes an LLM-powered framework to automatically

extract topics via topic modeling and generate hierarchical topics, which enables the identification of multi-level trends as well as the elucidation of the internal structure of fine-grained subtopics.

Topic modeling

Early probabilistic topic models—most notably Latent Dirichlet Allocation (LDA; Blei et al., 2003) and non-negative matrix factorization (NMF; Lee & Seung, 2000)—proved effective at uncovering latent thematic structures. However, their reliance on bag-of-words representations often degraded topic coherence and hindered interpretability. To overcome the flat structure of LDA, early hierarchical extensions were developed, most significantly the hierarchical Latent Dirichlet Allocation (hLDA; Blei et al., 2010). hLDA employs a non-parametric Bayesian approach to automatically infer a tree-like hierarchy from data. While mathematically rigorous, hLDA often faces challenges in semantic transparency, as the generated higher-level topics can be overly abstract and difficult for human experts to interpret (Mimno et al., 2011).

Recent developments address these limitations by integrating contextual embeddings and large language models (LLMs). For example, BERTopic (Grootendorst, 2022) combines transformer-based representations with clustering to produce semantically consistent topics, while prompt-based approaches (e.g., TopicGPT, GPTopic) improve interpretability by generating human-readable labels and explanations. Despite these advances, controlling topic granularity remains a persistent challenge (Mu et al., 2024).

In GPTopic (Reuter et al., 2024), users can refine granularity by interactively splitting or merging topics via prompt dialogues; yet this requires issuing prompts for each extracted topic individually, which is cumbersome. TopicGPT (Pham et al., 2023) treats the initially extracted topics as top-level categories and can derive more specific second-level subtopics from them; however, further hierarchical refinement beyond the second level (e.g., to third or fourth levels) is not supported.

Accordingly, this study adopts GPTopic to capitalize on the strengths of prompt-based topic modeling. By using GPTopic, the top 500 representative terms are leveraged to derive concise topic titles and descriptions that faithfully capture each topic's essence. To achieve controllable granularity, we further develop a prompt-based hierarchical structuring procedure that organizes topics into an explicit, multi-level taxonomy.

Scientific topic discovery

Scientific topic discovery is critical for synthesizing vast academic corpora, identifying emergent themes, and forecasting research trajectories in an era of information overload. Contemporary approaches increasingly leverage large language models (LLMs), network analysis, and hybrid methods to enhance interpretability, scalability, and predictive accuracy.

SciTopic (Liu et al., 2023) embeds metadata, titles, and abstracts into a textual encoder to create enriched document representations. A space-optimization module uses entropy-based sampling to target ambiguous points and LLM-guided triplet tasks to refine contrastive loss, delivering superior clustering and reducing semantic gaps for nuanced trend detection. Huang et al. (2024) build word co-occurrence graphs to find high-impact topics via centrality and extract structural topics via community detection, cutting reliance on heavy semantic models. This scales to large corpora and reveals partial alignment between structural and semantic themes for policy and research planning. Spielberger et al. (2025) present agentic RAG for topic discovery, storing external data as vectors and using an agent-based iterative loop where the LLM retrieves and refines queries. This improves efficiency, transparency, reproducibility, and reliability over prompt-only LLM methods.

These advancements signal a paradigm shift toward hybrid, AI-augmented topic discovery, integrating semantic, structural, and predictive methodologies. Accordingly, this study aims to

thoroughly discover and analyze scientific topics using a framework based on LLM-prompting for topic structuring, trend analysis, and network analysis.

Research method

Figure 1 illustrates the main components of our framework, including data acquisition, hierarchical topic induction, trend analysis, and network analysis for LLM-related domains and methods—linking descriptive dynamics with structural relationships to explain and forecast the evolution of LLM research.

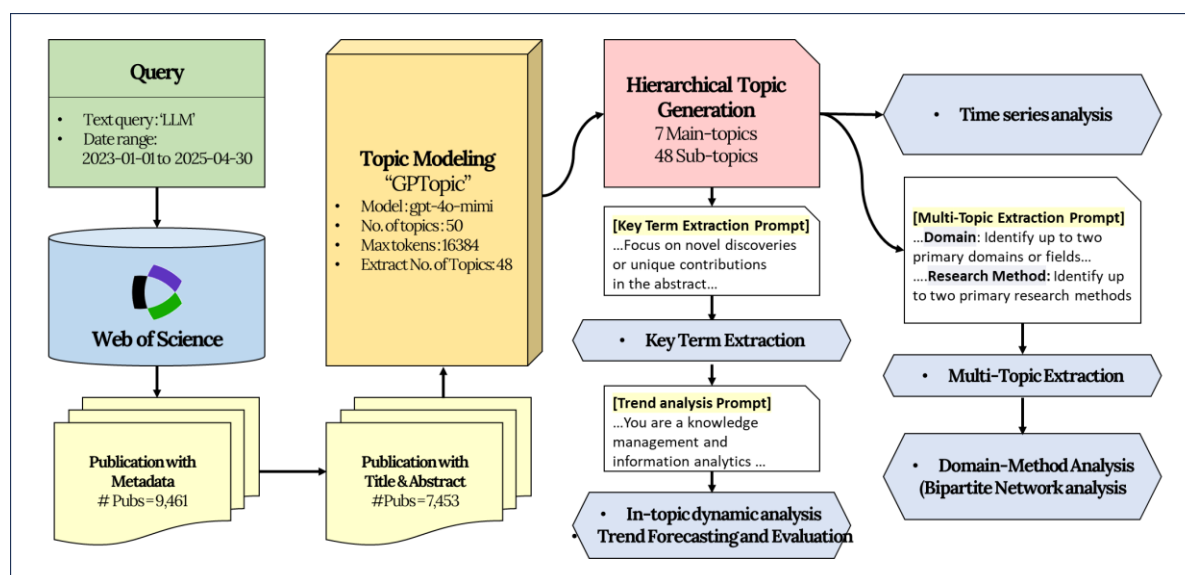


Figure 1. LLM-powered framework for research topic identification and structuring.

Data collection:

We queried the Web of Science Core Collection with the keyword ‘LLM’ to specifically capture the distinct surge in Large Language Model research while minimizing noise from broader terms like ‘generative AI’—which encompasses non-text modalities such as image and audio generation—or earlier technical precursors. We collected 9,461 records from January 2023—when LLM research surged—through April 2025. Among these, 7,453 publications containing both titles and abstracts were retained as the final dataset. The sample primarily comprises 3,424 journal articles and 4,029 conference papers. To capture temporal dynamics and field-specific shifts, the earliest online or formal publication date was retrieved for each record.

Hierarchical topic generation:

We applied topic modeling to the titles-and-abstracts corpus ($n = 7,453$) to extract research topics. Specifically, using GPTopic with the GPT-4o-mini model, we requested 50 topics and received 48. These topics are referred to as initial topics. We then prompted an LLM (GPT-4o) to generate main themes (hereafter referred to as macro-level topics) from these topics and to recover a hierarchical structure by splitting the macro-level topics into meso-level topics, which were further divided into micro-level topics. We define macro-level topics as broad, thematic areas; meso-level topics as coherent subdomains within each area; and micro-level topics as fine-grained, method- or case-specific clusters.

Research trend analysis:

With the hierarchical topics, we aggregate the publications by month to facilitate trend analysis. For each finer-grained topic that represents a more specific field, our framework examines in-topic dynamics by extracting key terms from publications in each topic by time. Specifically, we

aggregate documents in a topic by month and prompt an LLM to extract key terms and summarize key ideas from them. We present this analysis through a case study of topics in Medical Image and Data Processing. This topic is selected for its high publication volume in our corpus, which makes it a representative proxy for high-impact multimodal LLM research. By providing monthly summaries from January 2023 through April 2025, we obtain projections of post-April 2025 trends based on the temporal signals. The forecasting is evaluated with topics observed from May to August 2025.

Domain-method analysis:

Within each initial topic, we extract up to two domains and up to two research methods from the key terms of each publication to construct a domain-method bipartite network. Interactions between domains and research methods are then revealed by analyzing the network's coupling structure in terms of centrality and community organization.

Detailed structures of all prompts, including those used for keyword extraction, trend analysis, and domain-method labeling, are provided in Appendix A to allow for replication and adaptation of the framework.

Result

To validate the semantic coherence and interpretability of the 48 identified topics, we evaluated them using both statistical metrics and LLM-based qualitative assessments. Focusing on the top 20 domain-specific keywords, we computed the Normalized Pointwise Mutual Information (NPMI) based on word co-occurrences within a sliding window (Röder et al., 2015), obtaining a score of 0.45. This value indicates a solid level of thematic interpretability, suggesting that the keywords are semantically aligned rather than randomly distributed.

Furthermore, an LLM-as-a-judge assessment (detailed in the Appendix) evaluated the conceptual clarity of each topic on a 10-point scale, yielding a robust average score of 8.92. Notably, highly specialized domains—such as clinical diagnostics, cybersecurity, legal language processing, and autonomous driving systems—achieved perfect scores of 10/10, indicating that each topic represents a distinct and cohesive research trajectory.

Hierarchical topic generation

First, the results of the topic modeling are presented in Table 1. A total of 48 topics were extracted; among them, Topics 7 and 17, and Topics 12 and 40 were identified as duplicates and subsequently merged, as they appeared to be two semantically overlapping clusters that had been assigned the same label. We then performed hierarchical topic structuring, as shown in Figure 2. The topics were organized into seven macro-level themes, further grouped into meso- and micro-level subthemes, with each of the 46 extracted topics mapped to a corresponding micro-level subtheme. The seven macro themes are: (1) Artificial Intelligence and Machine Learning, (2) Educational Technology and Innovation, (3) Healthcare and Life Sciences, (4) Sentiment Analysis and Data Processing, (5) Robots and Automation, (6) Software and Security, and (7) Others.

In the first macro theme, the topics are split into 'language models' and 'machine learning technologies.' Each track moves from basics to optimization and evaluation, then to reliability, and finally to new advances. Security is a separate branch (1.3). In the second area, one branch focuses on using AI and conversational agents in education, and the other focuses on assessment and measurement. In the third area, medical AI divides into technology work (image/data processing) and clinical applications, while biomedical data is organized around analysis. The fourth area separates sentiment analysis into domain-specific work (like finance) and trust/verification, and includes a parallel strand on information retrieval, generation, and summarization. The fifth area covers robotics, with two main branches: human-robot interaction and autonomous systems (especially driving). The sixth area separates software development from security, covering the

whole software lifecycle. Finally, the seventh area groups other specialized applications—such as recommendation systems, geospatial data, energy modeling, multilingual machine translation, and legal language processing—under ‘Others’.

No.	Topic Label	Papers	No.	Topic Label	Papers
0	Large Language Models Research	2742	27	Hallucinations in Language Models	72
3	Medical AI Evaluation	334	33	Software Security Vulnerability Detection	71
29	Medical Language Models	326	34	Automated Code Generation and Testing	70
21	Multimodal Learning and Models	232	39	Geospatial Data Analysis	65
10	Robotic Language Learning and Planning	213	44	Energetic Materials and Decomposition	65
16	Efficient Machine Learning Systems	203	47	Autonomous Driving Systems	65
41	Educational Technology and AI	195	42	Emotion Recognition in AI	58
25	Medical Imaging and AI Evaluation	185	19	Machine Learning and Models	57
13	Cybersecurity and Language Models	179	7,17	Natural Language Processing Advances	52
5	Human-Robot Interaction and Learning	173	35	Mental Health Data Analysis	52
1	Educational Technology and Innovation	171	26	Educational Assessment and AI	51
4	Biomedical Data and Modeling	165	2	Sentiment Analysis in Finance	49
12,40	Artificial Intelligence in Healthcare	156	36	Muscle Growth and Obesity	46
28	Knowledge Graphs and Models	146	38	Bias in Language Models	45
30	Information Retrieval and Generation	139	15	Federated Learning and Model Optimization	44
6	Recommender Systems and Models	132	43	Multilingual Machine Translation	44
9	Reasoning and Language Models	120	46	Legal Language Processing	41
8	Conversational Agents in Gaming	114	23	Evaluation of Language Models	40
31	Prompt Engineering and Evaluation	102	22	Machine Learning Models and Evaluation	37
14	Software Engineering and Testing	101	37	Sustainable Energy Modeling	36
11	Fake News Detection	91	18	Language Model Optimization	21
32	Speech and Language Processing	73	24	Truthfulness and Reliability in AI	6
45	Chatbots in Education	73	20	Language Models and Techniques	1

Table 1. Initial topics generated by GPTopic, ranked by the number of publications, with duplicate topics merged.

Research trend analysis

Research trends at varying topic levels.

Monthly publication counts by macro-level theme are presented in Figure 3. For interpretability, Figure 3 employs two vertical axes: bars are plotted on the primary y-axis, while solid lines are plotted on the secondary y-axis. ‘Artificial Intelligence and Machine Learning’ dominates the overall volume, with a first peak in December 2023 and a cycle-high apex in August 2024 after a sharp mid-year surge. The December 2023 bump is consistent with concentrated outputs from EMNLP 2023 main (90), EMNLP Findings (76), and NeurIPS 2023 (63), while the August 2024 apex aligns with ACL 2024 long papers (188), short papers (12), and Findings (189), alongside KDD 2024 (34) and IJCAI 2024 (23). In contrast, ‘healthcare and life sciences’ exhibit a steady, structural uptrend rather than sporadic spikes, suggesting sustained diffusion of AI methods into clinical and biomedical workflows. Beyond these leading themes, ‘educational technology and innovation’ shows intermittent surges rather than a persistent climb, with localized peaks in April 2024 (66) and July 2024 (67) following a modest rise at year-end 2023 (18), and normalization into the low-to-mid 20s by late 2024. ‘Sentiment analysis and data processing’ remains thin through 2023 but scales in mid-2024, cresting in August 2024 (59) with elevated activity in June-July, before easing into the fourth quarter. ‘Robots and automation’ is event-driven, with sharp spikes in May 2024 (56) and August

2024 (58) and smaller rebounds thereafter. ‘Software and security’ transitions from low and volatile counts in 2023 to higher mid-2024 activity, peaking in June (30) and August (44) and tapering toward year-end. The residual ‘Others’ category remains small through 2023 but expands into a visible August 2024 peak (69), followed by a secondary bump in October (49) and subsequent moderation.

1. AI and Machine Learning		
• 1.1. Language Models		
1.1.1. Basic Research	0, 20	
1.1.2. Optimization and Evaluation	18, 23, 38	
1.1.3. Problems and Reliability	27, 24	
1.1.4. Advances and Technologies	7, 9, 28, 31	
• 1.2. Machine Learning Technologies		
1.2.1. Basic Technologies	16, 19, 22	
1.2.2. Advanced Technologies	15, 21	
• 1.3. Security and Reliability		
1.3.1. Cybersecurity	13	
2. Educational Technology and Innovation		
• 2.1. Educational Technology		
2.1.1. AI Utilization	41, 1	
2.1.2. Conversational Systems	45, 8	
• 2.2. Educational Assessment		
2.2.1. Evaluation Techniques	26	
3. Healthcare and Life Sciences		
• 3.1. Medical AI		
3.1.1. Image and Data Processing	25, 29	
3.1.2. AI Applications	12, 3	
• 3.2. Biomedical Data		
3.2.1. Data Analysis	4, 35, 36	
4. Sentiment Analysis and Data Processing		
• 4.1. Sentiment and Opinion Analysis		
4.1.1. Finance and Sentiment	2, 42	
4.1.2. Information Verification	11	
• 4.2. Information Retrieval and Generation		
4.2.1. Information Processing	30, 32	
5. Robots and Automation		
• 5.1. Human-Robot Interaction		
5.1.1. Learning and Interaction	5, 10	
• 5.2. Autonomous Systems		
5.2.1. Autonomous Driving Technologies	47	
6. Software and Security		
• 6.1. Software Development		
6.1.1. Development and Testing	14, 34	
• 6.2. Security		
6.2.1. Vulnerability Detection	33	
7. Others		
• 7.1. Recommendation Systems		
7.1.1. Models and Technologies	6	
• 7.2. Data Analysis		
7.2.1. Geospatial Data	39	
7.2.2. Energy Modeling	37, 44	
7.2.3. Multilingual Machine Translation	43	
7.2.4. Legal Language Processing	46	

Figure 2. Hierarchical structure of LLM-related research topics.

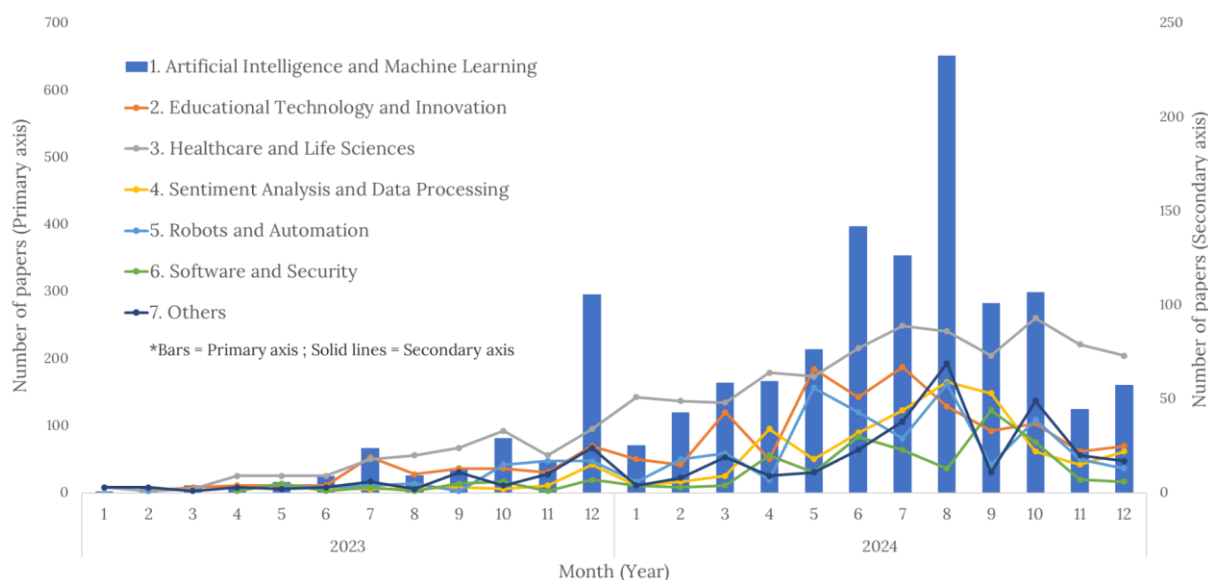


Figure 3. Monthly trends of macro topics from 2023 to 2024. Use of dual axes: bars are plotted on the primary y-axis, while solid lines are plotted on the secondary y-axis.

At a finer granularity, Figure 4 reports the monthly trajectories for the top five initial topics within ‘artificial intelligence’ and ‘machine learning’ (out of 17), to foreground the dominant signals while preserving readability and interpretability. Multimodal learning and models accounts for 232 papers, efficient machine learning systems for 203, cybersecurity and language models for 179, knowledge graphs and models for 146, and reasoning and language models for 120. Multimodal work accelerates markedly in mid-2024, reflecting momentum around text-image-audio integration and expanding evaluation suites (Sun et al., 2024; Zhang et al., 2024;). Meanwhile, efficiency-oriented research maintains a persistent climb as teams optimize memory, latency, and deployment costs across server and edge settings. Cybersecurity-focused LLM studies concentrate in mid-to-late summer, mirroring heightened attention to prompt injection, data poisoning, and model exfiltration threats. Knowledge graph integration and reasoning with LMs rise gradually, indicating growing interest in structured knowledge infusion and chain-based inference. Taken together, these patterns signal particularly strong and sustained interest in multimodal learning, systems efficiency, and security.

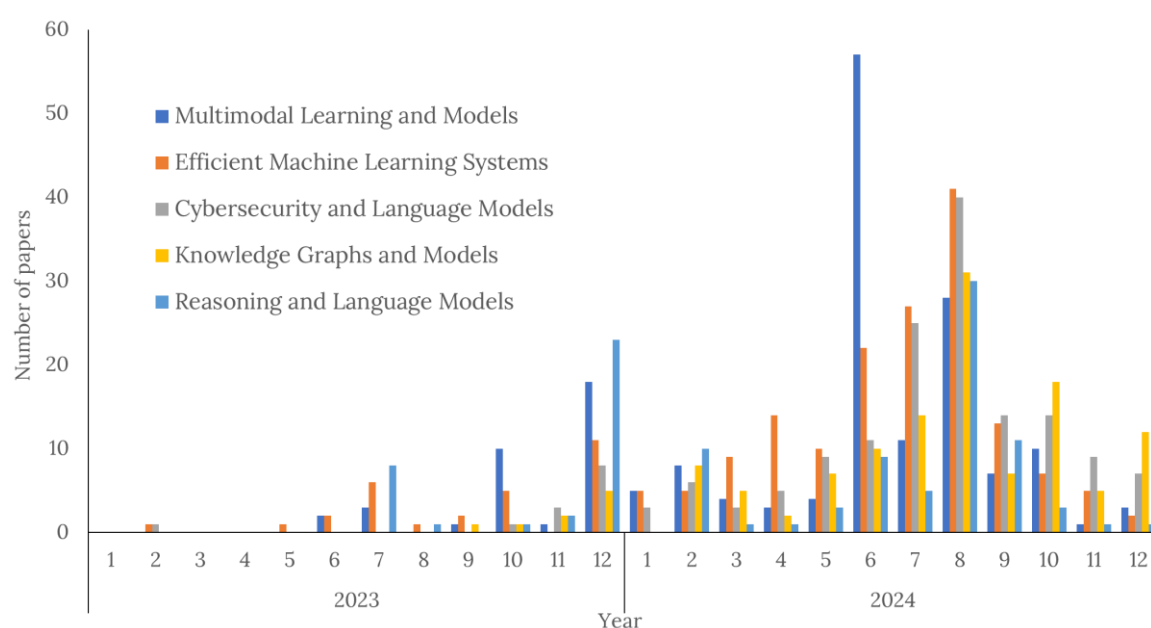


Figure 4. The monthly number of papers in the top 5 initial topics of AI and ML.

In-topic dynamic analysis

Time-series analysis has the advantage of quantitatively capturing trends but is limited in uncovering qualitative, fine-grained details. Therefore, we use a prompt-based approach to extract key terms from publications in selected topics and qualitatively characterize nuanced shifts in trends. Focusing on ‘healthcare and life sciences’ (the second most active area) and to examine multimodal research more closely, we analysed 511 papers within the medical AI subtopic of ‘image and data processing’. We first extracted monthly and yearly keyword insights and then utilized these insights to forecast subsequent research topics and evaluate the forecasts.

We extracted 20 key terms per abstract using GPT-4o with a keyword-extraction prompt. The prompt instructed the model to identify novel findings and unique contributions, outputting exactly 20 lowercase English terms as a comma-separated list. All outputs were validated for format compliance (count = 20, lowercase, comma-separated) before inclusion in Table 2. Building on these features, the monthly trend results derived using a trend-analysis prompt are presented in Figure 5, and the yearly trend results are shown in Figure 6.

2023: Early exploration and exploration of possibilities

2023 is the year when LLMs such as ChatGPT are introduced to the medical field, and the initial evaluation and applicability are explored.

- January

Analysis: ChatGPT's potential for medical applications is first noticed. Initial performance evaluations are conducted focusing on clinical reasoning and decision-making. **{Italian, Residency}** suggests its use in Italian resident education, and **{accuracy, performance}** focuses on testing the accuracy and performance of the model. This is the theoretical stage of exploring the medical potential of the LLM.

- April

Analysis: The scope of application of the LLM is expanded to **{pharmacists}**, **{Plastic Surgery}**, and **{dental medicine}**. **{Layered learning models}** suggests integration with educational models, and **{systematic review}** suggests systematic evaluation. Limitations and ethical issues such as **{misinformation, cybersecurity}** are also discussed for the first time. Practical applications (**telemedicine, clinical decision support**) are starting to take shape.

Figure 5. The monthly trend results were derived using a trend-analysis prompt.

Taken together, the monthly and yearly trajectories indicate a three-stage evolution. First, 2023 marks the nascent phase of LLM adoption in medicine. Early in the year, attention centers on feasibility and conceptual exploration—clinical reasoning, decision support, and resident education—followed by a mid-year broadening of use cases and the first systematic discussions of ethics and safety (e.g., misinformation, data governance, cybersecurity). From a yearly vantage point, medical education assessment and early diagnostic support emerge as the most active application areas in 2023, signaling low-risk, high-learning environments for initial deployment.

Second, in 2024, the center of gravity shifts from exploration to integration. Clinical workflows incorporate LLMs for electronic health record summarization, discharge documentation, and clinical trial matching, alongside specialty-specific pilots in fields such as neurosurgery and rheumatology. In parallel, evaluation practices mature: systematic reviews and shared benchmarks become more common, and metrics such as accuracy, F1, and AUROC see increasing standardization across studies. Methodologically, research tilts toward multimodal pipelines and domain-specialized models, reflecting the need to fuse imaging with text and to adapt general-purpose LLMs to clinical subdomains.

2023

- **Medical Education:** LLMs are tested in residency exams (e.g., Italian Residency, PSITE) and medical student assessments (e.g., community medicine).
- **Clinical Use:** Early diagnostic applications (e.g., ChatGPT in triage, rare disease diagnosis) and specialty-specific evaluations (e.g., otolaryngology, plastic surgery).
- **Challenges:** Misinformation and limitations in dental medicine and telemedicine highlight early hurdles.

2024

- **Clinical Integration:** LLMs support EHR summarization, discharge notes, and clinical trial matching. Specialties like neurosurgery and rheumatology see tailored applications.
- **Evaluation:** Systematic reviews and benchmarks (e.g., MedSumm, MedGPTEval) refine LLM assessment. Metrics like accuracy, F1-score, and AUC are standardized.
- **Innovation:** Multimodal models (e.g., CLIP, neuroGPT-X) and domain-specific LLMs (e.g., Zhongjing, CHIMED-GPT) address niche needs.

2025

- **Real-World Deployment:** LLMs enhance tumor boards, antibiotic prophylaxis, and telehealth (e.g., HIV care in Indonesia).
- **Specialization:** Models like Hypnos and Infective Endocarditis LLMs target specific fields. Fine-tuning for low-resource languages (e.g., Vietnamese) and TCM grows.
- **Ethical Focus:** Guidelines (e.g., TRIPOD-LLM) and equity-focused applications (e.g., social determinants of health) reflect maturing use.

Figure 6. The yearly trend results were derived using a trend-analysis prompt.

Micro level Topic	Initial Topic	Paper No.	Month	Key terms
Image and Data Processing	Medical Imaging and AI Evaluation	1	2023_1	chatgpt, clinical, reasoning, decision-making, medical, education, performance, accuracy, italian, residency, exam, results, research, potential, support, questions, analysis, cases, notional, participants
		2	2023_4	chatgpt, plastic, surgery, research, systematic, review, ideas, accuracy, novel, virtual, consultations, pre-operative, planning, patient, education, post-operative, care, utilization, community, software
		3	2023_4	chatgpt, large language models, dental medicine, implications, limitations, misinformation, telemedicine, clinical decision support, patient data confidentiality, cybersecurity, administrative personnel, academic writing, health information, text summarization, multilingual communication, efficient writing, risks, academic fields, benefits, tools
		4	2023_5	chatgpt, plastic surgery, psite, residency, accuracy, education, knowledge, information, logical, reasoning, learning, advancements, artificial intelligence, tool, evaluation, questions, coherence, clarifying, statistically significant, evidence-based medicine
		5	2023_5	chatgpt, otorhinolaryngology, consultants, medical, adequacy, coherence, conciseness, comprehensibility, case-based, evaluation, responses, ratings, technological, improvement, information, misinformation, semantic, answers, competence, study
		6	2023_6	chatgpt, otolaryngology, certification, examinations, performance, model, assessment, open-ended, accuracy, concordance, validity, consistency, hallucinations, responses, patient safety, results, study, trial, competency, method
		7	2023_6	chatgpt, community medicine, examination, medical students, learning, performance, study, scoring, responses, knowledge, relevancy, coherence, completeness, evaluators, india, undergraduate, self-directed learning, faculty oversight, pilot mode, medical education
		8	2023_6	diagnosis, triage, accuracy, chatgpt, physicians, ada, webmd, emergency, patients, clinical, data, recommendations, performance, emergent, results, analysis, unsafe, rates, matches, systems
	Medical Language Models	9	2023_4	layered, learning, models, pharmacists, student, ambulatory, care, preceptors, residents, clinical, practice, experience, skills, knowledge, development, preceptorship, effective, training, environment, insight
		10	2023_5	p-wave, catboost, lead, sinus, tachycardia, ecg, arrhythmias, detection, classification, accuracy, sensitivity, optimal, features, participants, recordings, machine, learning, bipolar, model, selection
		11	2023_6	best, practice, alerts, patients, screening, tool, dvt, anticoagulation, prophylaxis, model, biomed-roberta, accuracy, precision, recall, thromboembolism, health, care, ai, hipaa, improvement
		12	2023_7	diagnosis, rare, complex, language, models, education, performance, prompt, engineering, accuracy, bard, chatgpt-3.5, gpt-4, medalpaca, cases, support, human, majority, voting, research

Table 2. Examples of the extracted key terms

Third, by 2025, the emphasis moves to real-world deployment and specialization-at-scale. Representative use cases include tumor boards, antibiotic prophylaxis pathways, and telemedicine workflows, where reliability and turnaround-time matter as much as raw accuracy. Governance also advances: guidelines such as TRIPOD-LLM and fairness-centered protocols gain traction, encouraging transparent reporting, bias auditing, and equitable performance across patient subgroups. Overall, the pattern suggests a transition from ‘Can we apply LLMs?’ to ‘How do we sustain safe, efficient, and equitable clinical impact?’—with multimodality, domain adaptation, and standardized evaluation anchoring that shift.

Trend forecasting and evaluation

Prediction for May 2025 and Beyond

Based on the trajectory from 2023 to April 2025, here’s a forecast for trends post-May 2025:

Key Predicted Trends

Widespread Clinical Adoption:

LLMs will move beyond pilot studies to standard tools in clinical workflows, particularly in diagnostics (e.g., real-time decision support for emergencies) and patient management (e.g., personalized treatment plans). Example: Entries like {ChatGPT-4o, tumor boards, sarcoma} suggest multidisciplinary integration will expand to other complex diseases (e.g., oncology, cardiology).

Domain-Specific and Multimodal Models:

Specialized LLMs (e.g., Hypnos for anesthesiology, MedicalGLM for pediatrics) will proliferate, tailored to specific medical fields and regions (e.g., non-English languages like Vietnamese, Chinese). Multimodal capabilities (text, images, sensor data) will enhance diagnostics, as seen in {GPT-4, pathology, images} and {MAGDA, vision-language, diagnosis}.

Figure 7. Post-May 2025 research trajectories.

Using data from 2023 through April 2025, we generated a prompt-based forecast for post-May 2025 research trajectories; the synthesized results are presented in Figure 7. The forecast converges on two organizing pillars: (i) widespread clinical adoption—LLMs embedded in end-to-end workflows such as documentation, decision support, and care coordination—and (ii) domain-specific multimodal models that fuse imaging with structured and unstructured clinical text and are tuned for specialty-specific constraints.

To evaluate the plausibility of this forecast, we conducted an ex-post check using publications from May to August 2025. GPTopic was applied to 2,023 papers containing LLM-related keywords, yielding topic allocations at the article level. Within this corpus, 80 papers fell under ‘*medical imaging and diagnostics*’. We then screened their abstracts against two binary criteria—widespread clinical adoption and domain-specific & multimodal models—as summarized in Table 3.

Of these 80 papers (n=80), 26 (32.5%) satisfied both criteria (A), indicating concurrent signals of deployment and specialization. An additional 41 papers satisfied at least one criterion (B), bringing the total coverage (A+B) to 67 (83.75%). This high degree of alignment suggests that the forecast captured the dominant direction of research activity rather than isolated niches.

Criteria	Papers	Ratio (%)
A. Both Criteria	26	32.5
B. At least One Criterion	41	51.25
C. Neither	13	16.25
Total	80	100

Table 3. Validation of widespread clinical adoption with domain-specific and multimodal models

Two representative studies illustrate these trends. Zhu et al. (2025) introduced Woollie, an open-source, oncology-specialized LLM trained on oncology clinical data. Woollie achieved robust and generalizable performance—overall AUROC of 0.97 on MSK criteria (0.98 for pancreatic cancer) and 0.88 on UCSF external validation (0.95 for lung cancer)—underscoring the value of domain specialization with rigorous external validation. Agarwal et al. (2025) demonstrated hospital-specific domain adaptation for brain MRI classification using a BERT-based pipeline across two tertiary centers (KCH: 126,556 cases; GSTT: 86,032 cases), with statistically significant improvements over the baseline model, exemplifying practical steps toward institution-aware deployment.

The thirteen studies not included in the two main categories are characterized as follows. Seven of the studies emphasize the role of large language models (LLMs) in diverse radiological and oncological applications (Brown et al., 2025; Gallifant et al., 2025) such as automated radiology reading across multiple institutions (Moassefi et al., 2025), multicenter pulmonary disease prediction (Li et al., 2025), scientific hypothesis formation for breast cancer treatment (Abdel-Rehim et al., 2025), and evaluation of radiology residents' reports (Atsukawa et al., 2025). Two studies focus on machine learning prognostic models for pulmonary nodules (Wang et al., 2025) and endometrial cancer risk in a multiethnic cohort (Habeshian et al., 2025). Infectious disease and viral research are represented by three studies, highlighting improved COVID-19 severity assessment (de Carvalho et al., 2025), ultrasensitive respiratory virus detection (Cao et al., 2025), and recombination dynamics among canine and human influenza viruses (Kam et al., 2025). Additionally, one study explores LLM-based classification systems for oncological clinical trial information retrieval (Dennstadt et al., 2025), and another investigates lung lipid alterations related to smoking and e-cigarette use (McElroy et al., 2025).

From the standpoint of widespread clinical adoption, these papers show meaningful progress (multi-institutional validation, task-level accuracy, quality control for reporting, and potential impact on care pathways). Yet most remain in research or exploratory phases rather than fully embedded systems. Prospective, real-time workflow integration, regulatory readiness, and concrete patient or operational outcomes (e.g., time-to-decision, reduced unnecessary testing, cost-effectiveness) are still limited. From the domain-specific and multimodal models' perspective, many studies align well with specialized tasks and begin to combine data types (clinical variables, imaging, biomarkers). However, a substantial portion remains text-centric, constraining true multimodal fusion with pixel-level imaging, omics, or sensor data. Robustness across regions and languages and systematic fairness evaluations are also underdeveloped.

Domain-method analysis

We further conducted a domain-method analysis to uncover the fine-grained structure of documents that were initially classified via topic modeling. Because each document often mixes multiple themes, we selected the two most salient facets—domain and methodology—and applied LLM prompting to extract them. The resulting domain and methodology labels are summarized in Table 4. This dual-facet extraction enables more granular analysis by separating domain context from methodological intent. Building on these labels, we constructed a bipartite network between domains and methods (Figure 8) and computed node degrees for each side (Table 5).

No.	Domain1	Domain2	Methodology1	Methodology2
1	Emergency Medicine	Health Services	Classification Algorithms	LLM Evaluation
2	Health Communication	Global Health	LLM Fine-tuning	Dataset Development
3	Neurology	Epilepsy	Machine Learning Application	LLM Evaluation
4	Pediatrics	Medical Informatics	LLM Fine-tuning	Quality Evaluation
5	Neurology	Stroke	LLM Application	Prompt Engineering
6	Pharmacology	Medical Informatics	LLM Application	Terminology Mapping
7	Radiology	Medical Imaging	Systematic Review	Taxonomy
8	Patient Safety	Adverse Event Detection	LLM Evaluation	Framework Development
9	Surgery	Medical Education	LLM Evaluation	Exam Performance
10	Pharmacology	Drug Safety	Dataset Creation	LLM Evaluation

Table 4. The resulting domain and methodology of 511 papers under image and data processing within healthcare and life sciences.

As shown in Table 5, from the methodology perspective, two axes dominate: LLM evaluation and LLM application. Their high centrality indicates that empirical deployment and formal assessment are advancing in tandem rather than sequentially. On the domain side, medical informatics, medical education, and clinical informatics exhibit the highest degrees, suggesting that: (i) work spanning the broader medical data ecosystem is most active; (ii) the use and evaluation of LLMs in educational settings is rapidly scaling; and (iii) practical, workflow-oriented studies in clinical environments are gaining traction.

The bipartite graph in Figure 8 visually reinforces these patterns. Dense clusters form around LLM evaluation and LLM application, with frequent cross-links to medical informatics and clinical informatics; medical education connects strongly to both evaluation-oriented studies (e.g., exam performance, rubric design) and application-focused work (e.g., decision support training, simulation). High-degree methodological nodes such as comparative study, LLM fine-tuning, and framework development further indicate a maturing evidence base that blends head-to-head comparisons, domain adaptation, and standardized tooling. On the domain side, diagnostics, cardiology, emergency medicine, neurology, surgery, and radiology appear as secondary hubs, highlighting specialty-driven applied research that interfaces with imaging, triage, and procedural workflows.

No.	Domain	Degree	Method	Degree
1	Medical Informatics	98	LLM Evaluation	232
2	Medical Education	74	LLM Application	197
3	Clinical Informatics	54	Comparative Study	149
4	Diagnostics	41	LLM Fine-tuning	66
5	Cardiology	37	Review	57
6	Public Health	37	Framework Development	33
7	Emergency Medicine	35	Systematic Review	28
8	Neurology	33	Perspective	28
9	Surgery	27	LLM Development	27
10	Radiology	23	Information Retrieval	24

Table 5. Node degrees of domain and methods

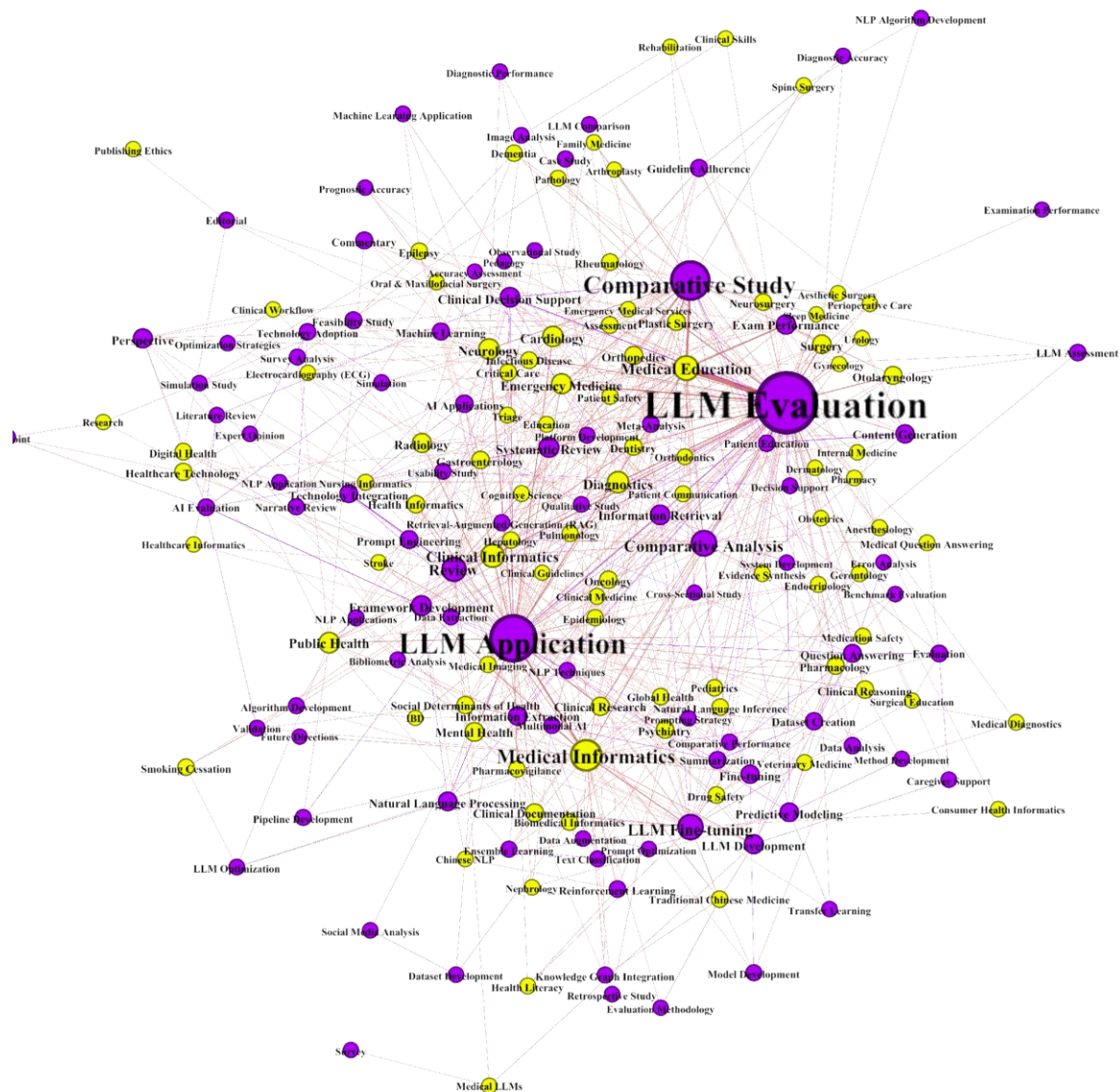


Figure 8. A bipartite network between domains and methods, with yellow nodes indicating **domains** and purple nodes indicating **methods**.

Taken together, these results point to a co-evolution dynamic: discussions of LLM application are tightly coupled with robust evaluation frameworks, and the most connected research arenas are those that span data governance, clinical workflow integration, and education. In practical terms, this means (a) broader informatics infrastructures are setting the pace for scalable adoption; (b) training and assessment in medical education are serving as low-risk testbeds for iterative improvement; and (c) clinically oriented studies are shifting from feasibility to reproducible, benchmarked deployment.

Implications

This study presents several significant implications for the future trajectory of LLM research. First, it highlights multimodality, systems efficiency, and cybersecurity as critical growth drivers in LLM-related research, thereby informing priority setting for technological advancement. Second, by identifying ‘*widespread clinical adoption*’ and ‘*domain-specific multimodal models*’ as two pivotal pillars for post-May 2025, the study provides strategic directions for researchers in medical LLM. Third, the observed co-evolution between methodological hubs (LLM evaluation and LLM application) and central anchors for knowledge diffusion (medical informatics, medical education, and clinical informatics) strongly underscores the imperative for close interdisciplinary collaboration. Finally, the hierarchical organization of topics, combined with the trend analysis and bipartite network methodologies employed in this study, is highly generalizable. Our framework can be adapted for topic discovery in other fields or for specific innovations (such as autonomous driving and quantum computing).

Limitations

This study offers valuable insights into the evolution of LLM research from 2023 to 2025; however, it is subject to certain limitations. One limitation is that the analysis spans from 2023 to August 2025. Given the rapid pace of LLM technology development, our reported results might not fully capture new paradigms or research trends that emerge beyond this specific period. Nonetheless, the proposed framework itself is capable of generating up-to-date results and can be applied to any specific period for which data are available. Another limitation is that the analysis was based on literature indexed in the Web of Science (WoS) and retrieved using the specific keyword ‘LLM’ to demonstrate our proposed framework. While this ensured a focused analysis of the core LLM literature, it may exclude relevant studies indexed solely under broader or related terms such as ‘*generative AI*,’ ‘*foundation models*,’ or ‘*transformer architectures*.’ This selection may omit significant research contributions not included in this corpus, such as conference proceedings and pre-print archives, which could bias the presented LLM research landscape. This limitation can be addressed by adjusting the scope of literature selection and expanding search queries, allowing researchers to define their own corpus when applying our framework for topic discovery.

Conclusion

This study proposes an LLM-powered framework to hierarchically generate research topics related to LLM from 2023 to 2025, enabling trend analysis at varying levels. Our results show that, in terms of macro-level topics (i.e., the most general themes), research in Artificial Intelligence and Machine Learning leads overall output with a pronounced August peak in 2024, while multimodal learning, systems efficiency, and cybersecurity emerge as primary growth drivers. In a case study of the Image and Data Processing topics, the monthly and yearly profiles reveal a three-stage transition: an introduction and exploration phase in 2023; workflow integration and the maturation of evaluation in 2024; and, by 2025, real-world deployment with increasing specialization alongside greater attention to equity and formal guidelines. Furthermore, using data up to April 2025, our framework effectively forecasts future developments in research topics. The two forecasted pillars, namely widespread clinical adoption and domain-specific multimodal models, cover 83.75% of publications in Medical Imaging and Diagnostics from May to August 2025. Finally, the domain-method bipartite network analysis reveals methodological and domain hubs, providing insights for interdisciplinary collaboration.

Acknowledgements

We are grateful to the College of Information Science at the University of Arizona for the research support provided for this work.

About the authors

Xuan Lu is an Assistant Professor in the College of Information Science at the University of Arizona, USA. She received her PhD from Peking University. Her research interests include human-centered data science, AI for Science, and the future of work. She can be contacted at luxuan@arizona.edu.

Youngsoo Ko is a PhD student in the College of Information Science at the University of Arizona, USA. He is currently pursuing his doctoral degree, and his research interests focus on natural language processing, large language models (LLMs), and multi-agent systems for scientific discovery. He can be contacted at youngsooko@arizona.edu.

References

- Abdel-Rehim, A., Zenil, H., Orhobor, O., Fisher, M., Collins, R. J., Bourne, E., ... & King, R. (2025). Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *Journal of the Royal Society Interface*, 22(227), 20240674. <https://doi.org/10.1098/rsif.2024.0674>
- Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M., & Dehghani, M. (2025). A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science*, 8(2), 25152459251325174. <https://doi.org/10.1177/25152459251325174>
- Agarwal, S., Wood, D., Murray, B. A., Wei, Y., Busaidi, A. A., Kafiabadi, S., ... & Booth, T. C. (2025). Impact of hospital-specific domain adaptation on BERT-based models to classify neuroradiology reports. *European Radiology*, 1-15. <https://doi.org/10.1007/s00330-025-11500-9>
- Angelov, D. (2020). Top2vec: Distributed representations of topics. arXiv preprint arXiv:2008.09470. <https://doi.org/10.48550/arXiv.2008.09470>
- Atsukawa, N., Tatekawa, H., Oura, T., Matsushita, S., Horiuchi, D., Takita, H., ... & Ueda, D. (2025). Evaluation of radiology residents' reporting skills using large language models: an observational study. *Japanese Journal of Radiology*, 1-9. <https://doi.org/10.1101/2024.11.06.24316838>
- Barman, K. G., Caron, S., Sullivan, E., de Regt, H. W., de Austri, R. R., Boon, M., ... & Weniger, C. (2025). Large physics models: Towards a collaborative approach with large language models and foundation models. arXiv preprint arXiv:2501.05382. <https://doi.org/10.48550/arXiv.2501.05382>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan), 993-1022. <https://doi.org/10.7551/mitpress/1120.003.0082>
- Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2), 1-30. <https://doi.org/10.1145/1667053.1667056>
- Brown, J. D., Lenchik, L., Doja, F., Kaviani, P., Judd, D., Probyn, L., ... & Retrouvey, M. (2025). Leveraging large language models in radiology research: a comprehensive user guide. *Academic Radiology*. <https://doi.org/10.1016/j.acra.2024.11.053>

- Cao, Y., Lau, P. N., Chin, A. W., He, Z., Au Yeung, C. C., Zeng, K., ... & Shum, H. C. (2025). A Phase Separation-Assisted Pre-Enrichment Method for Ultrasensitive Respiratory Virus Detection. *Advanced Science*, e06578. <https://doi.org/10.1002/advs.202506578>
- Cotfas, L. A., Sandu, A., Delcea, C., Diaconu, P., Frăsineanu, C., & Stănescu, A. (2025). From Transformers to ChatGPT: An Analysis of Large Language Models Research. *IEEE Access*. <https://doi.org/10.1109/access.2025.3600739>
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3), 1-45. <https://doi.org/10.1145/3641289>
- Dave, T., Athaluri, S. A., & Singh, S. (2023). ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in artificial intelligence*, 6, 1169595. <https://doi.org/10.3389/frai.2023.1169595>
- de Carvalho, S. C., Raboni, S. M., Bueno, L. B., Lapinski, B. A., Lissa, S. M., Amadeu, L. L. M., ... & Nogueira, M. B. (2025). Assessment of the relationship between hematologic parameters,(CPD), in screening for COVID-19 severity in women. *Future Science OA*, 11(1), 2540749. <https://doi.org/10.1080/20565623.2025.2540749>
- Dennstädt F, Windisch P, Filchenko I, Zink J, Putora PM, Shaheen A, Gaio R, Cihoric N, Wosny M, Aeppli S, Schmerder M, Shelan M, Hastings J. Application of a General Large Language Model-Based Classification System to Retrieve Information about Oncological Trials. *Oncology*. 2025 Jun 13:1-11. doi: 10.1159/000546946. Epub ahead of print. PMID: 40517775. <https://doi.org/10.1101/2024.12.03.24318390>
- Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P., Mavergames, C., & Gruen, R. L. (2014). Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, 11(2), e1001603. <https://doi.org/10.1371/journal.pmed.1001603>
- Gallifant, J., Chen, S., Jain, S. K., Moreira, P., Topaloglu, U., Aerts, H. J., ... & Bitterman, D. S. (2025). Reliability of large language model knowledge across brand and generic cancer drug names. *JCO Clinical Cancer Informatics*, 9, e2400257. <https://doi.org/10.1200/cci-24-00257>
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*. <https://doi.org/10.48550/arXiv.2203.05794>
- Habeshian, T. S., Park, S. Y., Conti, D., Wilkens, L. R., Le Marchand, L., & Setiawan, V. W. (2025). Inflammatory and insulinemic dietary and lifestyle patterns and incidence of endometrial cancer: the multiethnic cohort. *The American Journal of Clinical Nutrition*, 121(6), 1236-1245. <https://doi.org/10.1016/j.ajcnut.2025.04.020>
- Hofmann, T. (1999, August). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50-57). <https://doi.org/10.1145/312624.312649>
- Huang, Y., Liu, Q., Liu, J., & Hu, Y. (2022, September). Topic Discovery in Scientific Literature. In *CCF Conference on Computer Supported Cooperative Work and Social Computing* (pp. 481-491). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-2356-4_38
- Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, 28(12), 15873-15892. <https://doi.org/10.1007/s10639-023-11834-1>

- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873), 583-589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kabra, M., Nagpal, A., Sacheti, A., Kumar, M., & Joshi, S. (2024). GENWISE: Thematic Discovery from Textual Data. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning* (pp. 79-88).
- Kam, T. T., Bui, C. H., Yeung, H. W., Wong, P. C., Chin, A. W., Nicholls, J. M., ... & Chan, M. C. (2025). Viral characterization of the reassortants between canine influenza H3N2 and human pandemic (2009) H1N1 and avian H9N2 viruses in canine ex vivo tracheal explants. *Virology Journal*, 22, 218. <https://doi.org/10.1186/s12985-025-02836-1>
- Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13. doi/10.5555/3008751.3008829
- Li, P., Wang, Z., Zhang, X., Zhang, R., Jiang, L., Wang, P., & Zhou, Y. (2025). SciTopic: Enhancing Topic Discovery in Scientific Literature through Advanced LLM. *arXiv preprint arXiv:2508.20514*. <https://doi.org/10.48550/arXiv.2508.20514>
- Li, R., Mao, S., Zhu, C., Yang, Y., Tan, C., Li, L., ... & Yang, Y. (2025). Enhancing Pulmonary Disease Prediction Using Large Language Models With Feature Summarization and Hybrid Retrieval-Augmented Generation: Multicenter Methodological Study Based on Radiology Report. *Journal of Medical Internet Research*, 27, e72638. <https://doi.org/10.2196/72638>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>
- Maulana, F. I., Adi, P. D. P., & Widartha, V. P. (2024, November). A Comprehensive Review and Research Trends of Large Language Models (LLMs) in Artificial Intelligence. In *2024 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)* (pp. 346-351). IEEE. <https://doi.org/10.1109/icimcis63449.2024.10956533>
- McElroy, J. P., Song, M. A., Barr, J. R., Gardner, M. S., Kinnebrew, G., Kuklenyik, Z., ... & Shields, P. G. (2025). Lung lipids associated with smoking and ECIG use in a cross-sectional study and clinical trial. *Respiratory Research*, 26(1), 193. <https://doi.org/10.1186/s12931-025-03267-w>
- Mendoza-Revilla, J., Trop, E., Gonzalez, L., Roller, M., Dalla-Torre, H., de Almeida, B. P., ... & Lopez, M. (2024). A foundational large language model for edible plant genomes. *Communications Biology*, 7(1), 835. <https://doi.org/10.1038/s42003-024-06465-2>
- Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262-272). <https://doi.org/10.3115/1699571.1699627>
- Moassefi, M., Houshmand, S., Faghani, S., Chang, P. D., Sun, S. H., Khosravi, B., ... & Erickson, B. J. (2025). Cross-Institutional Evaluation of Large Language Models for Radiology Diagnosis Extraction: A Prompt-Engineering Perspective. *Journal of Imaging Informatics in Medicine*, 1-6. <https://doi.org/10.1007/s10278-025-01523-5>
- Mu, Y., Bai, P., Bontcheva, K., & Song, X. (2024b). Addressing Topic Granularity and Hallucination in Large Language Models for Topic Modelling. *arXiv preprint arXiv:2405.00611*. <https://doi.org/10.48550/arXiv.2405.00611>

- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2025). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5), 1-72. <https://doi.org/10.1145/3744746>
- Pan, H., Mudur, N., Taranto, W., Tikhonovskaya, M., Venugopalan, S., Bahri, Y., ... & Kim, E. A. (2025). Quantum many-body physics calculations with large language models. *Communications Physics*, 8(1), 49. <https://doi.org/10.1038/s42005-025-01956-y>
- Patil, R., & Gudivada, V. (2024). A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5), 2074. <https://doi.org/10.3390/app14052074>
- Pham, C., Hoyle, A., Sun, S., Resnik, P., & Iyyer, M. (2024). TopicGPT: A Prompt-based Topic Modeling Framework. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2956–2984. <https://doi.org/10.18653/v1/2024.naacl-long.164>
- Reuter, A., Thielmann, A., Weisser, C., Fischer, S., & Säfken, B. (2024). Gptopic: Dynamic and interactive topic representations. *arXiv preprint arXiv:2403.03628*. <https://doi.org/10.48550/arXiv.2403.03628>
- Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399–408). <https://doi.org/10.1145/2684822.2685324>
- Spielberger, G., Artinger, F. M., Reb, J., & Kerschreiter, R. (2025). Retrieval Augmented Generation for Topic Modeling in Organizational Research: An Introduction with Empirical Demonstration. *arXiv preprint arXiv:2502.20963*. <https://doi.org/10.48550/arXiv.2502.20963>
- Sun, K., Bai, Y., Qi, J., Hou, L., & Li, J. (2024). MM-MATH: Advancing Multimodal Math Evaluation with Process Evaluation and Fine-grained Classification. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1358–1375. <https://doi.org/10.18653/v1/2024.findings-emnlp.73>
- Wang, D., Qiu, J., Li, R., & Tian, H. (2025). Development and interpretation of machine learning-based prognostic models for predicting high-risk prognostic pathological components in pulmonary nodules: integrating clinical features, serum tumor marker and imaging features. *Journal of Cancer Research and Clinical Oncology*, 151(6), 190. <https://doi.org/10.1007/s00432-025-06241-7>
- Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., ... & Han, X. (2024). AI for social science and social science of AI: A survey. *Information Processing & Management*, 61(3), 103665. <https://doi.org/10.1016/j.ipm.2024.103665>
- Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024). MM-LLMs: Recent Advances in MultiModal Large Language Models. *Findings of the Association for Computational Linguistics ACL 2024*, 12401–12430. <https://doi.org/10.18653/v1/2024.findings-acl.738>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2). <https://doi.org/10.48550/arXiv.2303.18223>
- Zhu, M., Lin, H., Jiang, J., Jinia, A. J., Jee, J., Pichotta, K., Waters, M., Rose, D., Schultz, N., Chalise, S., Valleru, L., Morin, O., Moran, J., Deasy, J. O., Pilai, S., Nichols, C., Riely, G., Braunstein, L.

Z., & Li, A. (2025). Large language model trained on clinical oncology data predicts cancer progression. *Npj Digital Medicine*, 8(1). <https://doi.org/10.1038/s41746-025-01780-2>

© [CC-BY-NC 4.0](#) The Author(s). For more information, see our [Open Access Policy](#).

Appendix A. LLM prompts for framework implementation

A.1. Key term extraction prompt. The following prompt was utilized to identify and extract the most significant research elements from each publication abstract:

System role: You are an expert linguistic analyst specializing in text analysis and key term extraction. **Task:** Analyze the provided abstract, identify novel findings or newly discovered elements, and extract the 20 most significant English words or phrases that represent the core of these findings. Ensure the terms are concise, relevant, and reflect the innovative aspects of the abstract. **Constraints:**

- Format: {term1}, {term2}, ..., {term20} (comma-separated list).
- Focus on unique contributions and novel discoveries.
- Prioritize specific, impactful terms over generic context.
- All terms must be in English and lowercase for consistency.

A.2. Trend interpretation and prediction prompt. The following prompt was used to synthesize the monthly and yearly progression of research topics into a structured narrative:

Persona: You are a knowledge management and information analytics professional with an exceptional ability to systematically analyze and explain complex data and the evolution of knowledge. **Role:** Your role is to track the evolution of knowledge in detail by year and month, based on a given topic or dataset, and to explain it in a logical and structured manner. **Task:** Describe the evolution of knowledge in detail by year and month, based on the provided [In-topic Keyword Data]. **Guidelines:** > * Maintain scholarly rigor and clarity.

- Tailor the description to be accessible to specialists while using structured, logical language.
- Focus on identifying specific shifts in research focus and the accumulation of information over time.

A.3. Domain-method extraction prompt. The following prompt was used to systematically categorize each publication into its respective domain and methodology:

System role: You are a highly skilled research assistant tasked with analyzing a research paper. **Task:** Extract the following information with precision and clarity:

1. **Domain:** Identify up to two primary domains or fields of study (e.g., ‘machine learning’, ‘biomedicine’).
2. **Research method:** Identify up to two primary research methods or approaches (e.g., ‘experimental’, ‘qualitative analysis’). **Constraints:**
 - Extract information based solely on the provided content.
 - Prioritize the most prominent elements if multiple are present.

- Output must be in a structured **JSON format**: `{ "domains": ["domain1", "domain2"], "research methods": ["method1", "method2"] }`

Appendix B. Topic coherence by LLM-as-a-judge assessment

B.1. Summary of topic coherence score by LLM-as-a-judge

No.	Topic Label	Score	No.	Topic Label	Score
0	Large Language Models Research	6	24	Truthfulness and Reliability in AI	9
1	Educational Technology and Innovation	8	25	Medical Imaging and AI Evaluation	10
2	Sentiment Analysis in Finance	9	26	Educational Assessment and AI	8
3	Medical AI Evaluation	10	27	Hallucinations in Language Models	9
4	Biomedical Data and Modeling	10	28	Knowledge Graphs and Models	9
5	Human-Robot Interaction and Learning	9	29	Medical Language Models	10
6	Recommender Systems and Models	10	30	Information Retrieval and Generation	9
7	Natural Language Processing Advances	7	31	Prompt Engineering and Evaluation	8
8	Conversational Agents in Gaming	8	32	Speech and Language Processing	10
9	Reasoning and Language Models	9	33	Software Security Vulnerability Detection	10
10	Robotic Language Learning and Planning	10	34	Automated Code Generation and Testing	10
11	Fake News Detection	9	35	Mental Health Data Analysis	10
12	Artificial Intelligence in Healthcare	8	36	Muscle Growth and Obesity	10
13	Cybersecurity and Language Models	9	37	Sustainable Energy Modeling	8
14	Software Engineering and Testing	10	38	Bias in Language Models	9
15	Federated Learning and Model Optimization	9	39	Geospatial Data Analysis	8
16	Efficient Machine Learning Systems	9	40	Artificial Intelligence in Healthcare	8
17	Natural Language Processing Advances	8	41	Educational Technology and AI	8
18	Language Model Optimization	8	42	Emotion Recognition in AI	10
19	Machine Learning and Models	9	43	Multilingual Machine Translation	9
20	Language Models and Techniques	6	44	Energetic Materials and Decomposition	10
21	Multimodal Learning and Models	10	45	Chatbots in Education	8
22	Machine Learning Models and Evaluation	9	46	Legal Language Processing	10

B.2. In-depth analysis of high-scoring specialized domains

The reason for the perfect coherence score (10/10) in these specialized fields is the use of domain-specific terminology. This terminology has very little overlap with general large language model (LLM) methodologies. This exclusivity creates a closed semantic loop, guaranteeing high topic coherence.

1. Clinical diagnostics and healthcare

- **Topic 3 (Medical AI evaluation)**
 - **Keywords:** *clinical, reports, radiology, diagnostic, diagnosis, report*
 - **Rationale:** This topic forms a highly cohesive cluster focused on the interpretation of medical imaging. The strong semantic link between ‘*radiology*’ and ‘*diagnostic reports*’ reflects a specific clinical practice rather than general text generation, ensuring no overlap with non-medical domains.
- **Topic 25 (Medical imaging and AI evaluation)**

- **Keywords:** *medical, clinical, questions, patients, diagnosis, scores*
- **Rationale:** This represents the research trajectory of evaluating LLM accuracy in clinical reasoning. The coexistence of ‘*questions*’ and ‘*diagnosis*’ indicates a clear and distinct focus on medical decision support systems (DSS).
- **Topic 29 (Medical language models)**
 - **Keywords:** *medical, clinical, healthcare, patients, notes*
 - **Rationale:** Centered on "clinical notes," this topic accurately captures the specialized task of processing Electronic Health Records (EHR), a well-defined sub-domain within healthcare informatics.

2. Cybersecurity and software integrity

- **Topic 13 (Cybersecurity and language models)**
 - **Keywords:** *privacy, attacks, security, adversarial, malicious, jailbreak, cybersecurity*
 - **Rationale:** The term ‘*jailbreak*’ combined with ‘*adversarial*’ and ‘*malicious*’ identifies a very narrow and specialized research path—AI safety and offensive security—which is clearly distinguished from general software development or conversational AI.
- **Topic 33 (Software security vulnerability detection)**
 - **Keywords:** *vulnerabilities, detection, software, security, source, static, fuzzing*
 - **Rationale:** The inclusion of highly technical terms like ‘*fuzzing*’ (an automated software testing technique) and ‘*static analysis*’ proves that the framework successfully isolated expert-level security research from broader IT topics, achieving maximum conceptual clarity.

3. Legal language processing

- **Topic 46 (Legal language processing):**
 - **Keywords:** *legal, law, domain, reasoning, case, cases, documents.*
 - **Rationale:** The concentration of terms such as ‘*case law*,’ ‘*legal documents*,’ and ‘*legal reasoning*’ demonstrates a distinct trajectory focused on the formal structures of the judicial system. Unlike general reasoning tasks, this topic is anchored in the specific linguistic and structural requirements of the legal profession.

4. Autonomous driving systems

- **Topic 47 (Autonomous driving systems):**
 - **Keywords:** *driving, traffic, autonomous, scenarios, control, safety, vehicles.*
 - **Rationale:** This topic is defined by ‘*traffic scenarios*,’ ‘*autonomous control*,’ and ‘*vehicle safety*.’ It represents the intersection of LLMs with physical-world robotics and real-time decision-making, forming a semantically isolated group that focuses on the safety-critical aspects of transportation.